If an error occurs during the reading out of the read documents, the read document is displayed on a display screen and the data can be read out only by marking corresponding fields in the read document. Here, if it is required, additional master documents are automatically produced on the basis of the marked read documents, or existing master documents are correspondingly corrected. This system is easy enough to use that no special computer or software knowledge is necessary.

A method that supports an operator in the generation of electronic templates for a form recognition system arises from US 5,317,646. For this, a form not provided with data (what is known as a master form) is shown on a screen, and the user can identify the data fields with a pointer device. The coordinates that bound the corresponding region are automatically detected after which a single point within this region has been selected by the operator. Templates for the automatic form recognition can be created simply and quickly with this method.

In Casey R.G. et al., "Intelligent Forms Processing", IBM Systems Journal volume. 29 (1990) Nr. 3, pages 435 through 450, a form recognition method is described in which a scanned-in form is analyzed by means of image processing techniques and is compared with other stored template forms. In the event that no correlation with a template form is found, a new template form must be generated via input on a computer. In the generation of a template, the scanned form is shown on the screen and the boundary lines of the input fields are marked with a pointer device.

A two-stage method in which form templates can be initially input and documents can be automatically read out using the input form templates arises from US 2002/141660 A1. Form templates to be input are scanned, and the operator indicates input fields with a cursor. The position and size of the input fields is stored. The operator can also determine the data type associated with each data field. Given automatic reading of forms, these are scanned in and automatically read out using the data fields contained in the stored form documents. In the event that an error occurs in the readout, the operator can correct the errors via the keyboard.

US 6,028,970 concerns a method and a system for automatic text recognition (OCR). The system comprises an error correction module ("error correction logic module"). This error correction module is applied to clearly detectable data errors in order to correct these. These corrections are executed automatically. Not only errors of individual letters are hereby detected, but rather errors in context are analyzed and correspondingly corrected. An error that cannot be automatically corrected can be communicated to the operator by means of an error message. The operator can then assess and, if applicable, correct the text generated by means of the text recognition.

The present invention is based on the object of creating a method and a system for acquiring data from machine-readable documents in which the inputting of the data is significantly simplified in comparison with the known methods in cases in which data cannot be automatically extracted.

This object is achieved by a method having the features of Claim 1 and by a system having the feature of Claim 16. Advantageous constructions of the present invention are indicated in the respective subclaims.

With the methods explained above, data can be acquired from a plurality of machine-readable documents, the data being assigned to a database in that individual data are extracted from the document as automatically as possible and are entered into corresponding database fields. If data cannot be extracted with the necessary degree of reliability for one or more particular database fields of a document, for example because an error has been determined, caused for example by the fact that no data or false data are present in the document at the point at which the data are to be read, or that during the reading in of this document using an OCR method one or more characters are falsely converted, then according to the present invention the following steps are executed:

- displaying of the document on a display screen,
- indication on the display screen of the database field for which the data cannot be extracted with the necessary degree of reliability,

- execution of a proposal routine with which string sections in the vicinity of a pointer on the display screen that can be moved by a user are selected, marked, and proposed for extraction.

Claims

1. Method for acquiring data from machine-readable documents, the data being assigned to a database, by extracting individual data from the document as automatically as possible and entering them into corresponding database fields, and, if data cannot be extracted from a document with the required degree of reliability for one or more particular database fields, executing the following steps:

- displaying of the document on a display screen,
- indication on the display screen of the database field for which the data cannot be extracted with the necessary degree of reliability,
- execution of a proposal routine with which string sections in the vicinity of a pointer on the display screen that can be moved by a user are selected, marked, and proposed for extraction.

2. Method according to Claim 1,
characterized in that the string section is selected, marked, and proposed for extraction in accordance with concept information assigned to the database field.

3. Method according to Claim 2,
characterized in that the concept information describes the syntax and/or the semantics of the database field, so that the proposal routine selects and marks a string section that is to be marked in a manner corresponding to the syntax or to the semantics of the respective database field.

4. Method according to Claim 3,
characterized in that the information concerning syntax describes the number of numerals and/or letters and/or predetermined formats of the string section that is to be read.

5. Method according to Claim 3 or 4,
characterized in that
the information concerning semantics describes specified terms, for example using a lexicon.

6. Method according to one of Claims 1 to 5,

characterized in that a string section is selected that is situated between two limiting **[or: boundary]** characters.

7. Method according to Claim 6,
characterized in that the limiting characters include empty characters and/or punctuation marks.

8. Method according to one of Claims 1 to 7,
characterized in that the text of documents in graphic representation is first converted into coded text using an OCR method, and the proposal routine represents, in addition to the marked string section in graphic representation, the coded text of this string section.

9. Method according to one of Claims 1 to 7,
characterized in that in addition to the marked string section, this string section is displayed again on the display screen in an enlarged representation.

10. Method according to one of Claims 1 to 9,
characterized in that after the marking of a string section, the proposal routine activates a function with which the content of the marked string section is transferred into the database through the actuation of one or more predetermined keys.

11. Method according to one of Claims 1 to 10,
characterized in that during the execution of the proposal routine, after the movement of the pointer a predetermined time wait interval is observed, during which the pointer must not be moved, before a string section is selected.

12. Method for acquiring data from machine-readable documents, the data being assigned to a database, in particular according to one of Claims 1-11,

characterized in that after data have been read from a first table row into corresponding database fields, the further table entries are automatically determined through a comparison of string sections situated under the first table row with the string sections of the first table